PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20311
doi: 10.3897/tdwgproceedings.1.20311

OPEN ACCESS

Conference Abstract

Towards a comprehensive workflow for biodiversity data in R

Tomer Gueta‡, Vijay Barve§, Ashwin Agrawal|, Thiloshon Nagarajah¶, Yohay Carmel‡

‡ Department of Civil and Environmental Engineering, The Technion – Israel Institute of Technology, Haifa, Israel
§ Florida Museum of Natural History, Gainesville, United States of America
| Indian Institute Of Technology(IIT) -BHU, Varanasi, India
¶ Informatics Institute of Technology, Colombo, Sri Lanka

Corresponding author: Tomer Gueta (tomer.gu@gmail.com), Vijay Barve (vijay.barve@gmail.com)

Received: 15 Aug 2017 | Published: 15 Aug 2017

Citation: Gueta T, Barve V, Agrawal A, Nagarajah T, Carmel Y (2017) Towards a comprehensive workflow for biodiversity data in R. Proceedings of TDWG 1: e20311. https://doi.org/10.3897/tdwgproceedings.1.20311

Abstract

Increasing number of scientists are using R for their data analyses, however, proficiency required to manage biodiversity data in R is considerably rarer. Since, users need to retrieve, manage and assess high-volume data with inherent complex structure (Darwin Core standard, DwC), various R packages dealing with biodiversity data and specifically data cleaning have been published. Though numerous new procedures are now available, implementing them require users to provide a great deal of efforts in exploring and learning each R package. For the common users, this task can be daunting. In order to truly facilitate data cleaning using R, there is an urgent need for a package that will fully integrate functionality of existing packages, enhance their functionality, and simplify its implementation. Furthermore, it is also necessary to identify and develop missing crucial functionalities.

We are attempting to address these issues by developing two projects under Google Summer of Code (GSoC)-- an international annual program that matches up students with open source organizations to develop code during their summer break. The first project is dealing with the integration challenge by developing a taxonomic cleaning workflow; standardizing various spatial and temporal data quality checks; and enhancing different data retrieval and data management techniques. The second project aims at advancing new and exciting features, such as establishing a flagging system (HashMap-like) in R, an

© Gueta T et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

innovative set of DwC summary tables, and developing new techniques for outliers analysis.

The products of these projects lay down crucial infrastructure for data quality assessment in R. Obviously this is a work in progress and needs further inputs. By developing a comprehensive framework for handling biodiversity data, we can fully harness the synergetic quality of R, and hopefully supply more holistic and agile solutions for the user.

Keywords

R, user-level, data quality assessment, Google Summer of Code, biodiversity data

Presenting author

Tomer Gueta and Vijay Barve

Grant title

This research is supported by the Israel Science Foundation (ISF) and the Google Summer of Code program